Homework Assignment 2

IT 531, Summer 2018

**Question 1**

Complete following example where PlayTennis is a class attribute. Construct a decision tree using ID3 algorithm. For each node, show the entropy/information in the data set and the potential gain for each possible attribute. Also, show the final tree.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Day** | **Outlook** | **Temperature** | **Humidity** | **Wind** | **PlayTennis** |
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

Answer :

1. First level

|  |  |
| --- | --- |
| First level | Info(D) = I(9,5) = 0.94  Info(D) of outlook = 0.694 Gain(outlook) = 0.246  Info(D) of tempature = 0.911 Gain(tempature) = 0.029  Info(D) of humidity= 0.788 Gain(humidity) = ~0.15  Info(D) of wind= 0.892 Gain(wind) = ~0.048 |
| node | Choose outlook as first node |

1. Second level

Outlook = sunny

|  |  |  |  |
| --- | --- | --- | --- |
| **Temperature** | **Humidity** | **Wind** | **PlayTennis** |
| Hot | High | Weak | No |
| Hot | High | Strong | No |
| Mild | High | Weak | No |
| Cool | Normal | Weak | Yes |
| Mild | Normal | Strong | Yes |

Outlook = rain

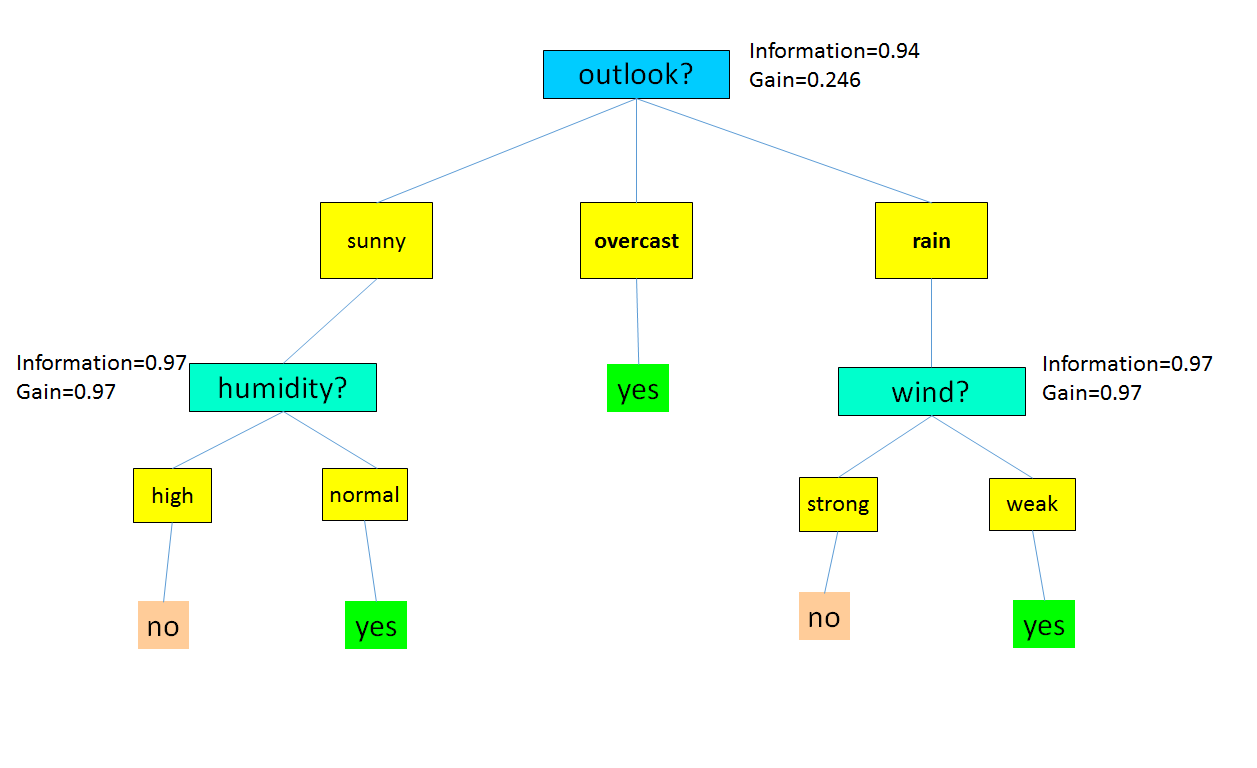
|  |  |  |  |
| --- | --- | --- | --- |
| **Temperature** | **Humidity** | **Wind** | **PlayTennis** |
| Mild | High | Weak | Yes |
| Cool | Normal | Weak | Yes |
| Cool | Normal | Strong | No |
| Mild | Normal | Weak | Yes |
| Mild | High | Strong | No |

|  |  |
| --- | --- |
| Outlook = sunny | Info(D) = I(2,3) = 0.97  Info(D) of tempature =0.4 Gain(tempature) = 0.57  Info(D) of humidity=0 Gain(humidity) = 0.97  Choose humidity as second node |
| Outlook = overcast | Pure , all play\_tennis is Yes |
| Outlook =rain | Info(D) = I(3,2) = 0.97  Info(D) of tempature =0.95 Gain(tempature) =0.019  Info(D) of humidity=0.95 Gain(humidity) = 0.019  Info(D) of wind=0 Gain(wind) = 0.97  Choose wind as second node |

1. Third level

|  |  |
| --- | --- |
| Outlook = sunny  Humidity = high | Pure , all play\_tennis is No |
| Outlook = sunny  Humidity = normal | Pure , all play\_tennis is Yes |

|  |  |
| --- | --- |
| Outlook = rain  Wind = strong | Pure , all play\_tennis is No |
| Outlook = rain  Wind = weak | Pure , all play\_tennis is Yes |

****

**Question 2**

Suppose you have the following training set with three boolean input x, y and z, and a boolean output U.

|  |  |  |  |
| --- | --- | --- | --- |
| **x** | **y** | **z** | **Class** |
| 1 | 0 | 0 | 0 |
| 0 | 1 | 1 | 0 |
| 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 |
| 0 | 0 | 1 | 1 |
| 0 | 1 | 0 | 1 |
| 1 | 1 | 0 | 1 |

Using the Naive Bayes algorithm, how would the following examples be classified?  Be sure to show your work as well as the final classification.

1. <x=1, y=1, z=1>
2. <x=0, y=1, z=0>

Answer:

a. class = 1

|  |  |
| --- | --- |
| Using Bayes therem | P(C|X) = P(X|C) \* P(C) /P(X) |
| Calculate the  P(class=0| x=1, y=1, z=1) | 1. P(X|C) = P(X1|C) \* P(X2|C)\*…\*P(Xn|C)   = P(x=1|class =0)  \* P(y=1|class=0)  \* P(z=1|class=0)  =1/3 \* 1/3 \* 2/3~=0.074   1. P(C) = P(class=0) = 3/7 ~=0.43 2. P(X) = P(X1)\*P(X2)\*…\*P(Xn)   = P(x=1)  \* P(y=1)  \* P(z=1)  = 3/7 \* 3/7 \* 3/7 ~=0.078  P(C|X) ~= 0.074\*0.43 / 0.078 ~=0.4 |
| Calculate the  P(class=1| x=1, y=1, z=1) | 1. P(X|C) = P(X1|C) \* P(X2|C)\*…\*P(Xn|C)   = P(x=1|class =1)  \* P(y=1|class=1)  \* P(z=1|class=1)  =2/4 \* 2/4 \* 1/4~=0.0625   1. P(C) = 4/7 ~=0.57 2. P(X) = P(X1)\*P(X2)\*…\*P(Xn)   = P(x=1)  \* P(y=1)  \* P(z=1)  = 3/7 \* 3/7 \* 3/7 ~=0.078  P(C|X) ~= 0.0625\*0.57 / 0.078 ~=0.45 |
| Compare P(class=0| x=1, y=1, z=1) and P(class=1| x=1, y=1, z=1) | P(class=1| x=1, y=1, z=1) is larger |

1. class = 1

|  |  |
| --- | --- |
| Using Bayes therem | P(C|X) = P(X|C) \* P(C) /P(X) |
| Calculate the  P(class=0| x=0, y=1, z=0) | 1. P(X|C) = P(X1|C) \* P(X2|C)\*…\*P(Xn|C)   = P(x=0|class =0)  \* P(y=1|class=0)  \* P(z=0|class=0)  =2/3 \* 1/3 \* 1/3~=0.074   1. P(C) = P(class=0) = 3/7 ~=0.43 2. P(X) = P(X1)\*P(X2)\*…\*P(Xn)   = P(x=0)  \* P(y=1)  \* P(z=0)  = 4/7 \* 3/7 \* 4/7 ~=0.245  P(C|X) ~= 0.074\*0.43 / 0.245 ~=0.13 |
| Calculate the  P(class=1| x=0, y=1, z=0) | 1. P(X|C) = P(X1|C) \* P(X2|C)\*…\*P(Xn|C)   = P(x=0|class =1)  \* P(y=1|class=1)  \* P(z=0|class=1)  =2/4 \* 2/4 \* 3/4~=0.1875   1. P(C) = 4/7 ~=0.57 2. P(X) = P(X1)\*P(X2)\*…\*P(Xn)   = P(x=0)  \* P(y=1)  \* P(z=0)  = 4/7 \* 3/7 \* 4/7 ~=0.245  P(C|X) ~= 0.1875\*0.57 / 0.245 ~=0.43 |
| Compare P(class=0| x=0, y=1, z=0) and P(class=1| x=0, y=1, z=0) | P(class=1| x=0, y=1, z=0) is larger |

**Question 3**

The following table consists of training data from an employee database. The data have been generalized; for example, “31…35" for age represents the age range of 31 to 35. The class label is status.

|  |  |  |  |
| --- | --- | --- | --- |
| Department | Age | Salary | Status |
| Sales | 31…35 | 46K…50K | Senior |
| Sales | 26…30 | 26K…30K | Junior |
| Sales | 31…35 | 31K…35K | Junior |
| Systems | 21…25 | 46K…50K | Junior |
| Systems | 31…35 | 66K…70K | Senior |
| Systems | 26…30 | 46K…50K | Junior |
| Systems | 41…45 | 66K…70K | Senior |
| Marketing | 36…40 | 46K…50K | Senior |
| Marketing | 31…35 | 41K…45K | Junior |
| Secretary | 46…50 | 36K…40K | Senior |
| Secretary | 26…30 | 26K…30K | Junior |

1. Use ID3 algorithm, which selects attribute with highest information gain to construct a decision tree from the given data.
2. Given a data tuple having the values “systems", “26. . . 30", and “46K…50K" for the attributes department, age , and salary , respectively, what would a Naive Bayes classification of the status for the tuple be? Note: **Use Laplacian correction**

Answer :

1. decision tree
2. First level

|  |  |
| --- | --- |
| First level | Info(D) = I(5,6) = 0.994  Info(D) of department =0.977 Gain(department) = 0.012  Info(D) of age = 0.363 Gain(age) = 0.626  Info(D) of salary= 0.363 Gain(salary) = 0.626 |
| node | Choose age as first node |

1. Second level

age =21~25

|  |  |  |
| --- | --- | --- |
| Department | Salary | Status |
| Systems | 46K…50K | Junior |

age =26~30

|  |  |  |
| --- | --- | --- |
| Department | Salary | Status |
| Sales | 26K…30K | Junior |
| Systems | 46K…50K | Junior |
| Secretary | 26K…30K | Junior |

age =31~35

|  |  |  |
| --- | --- | --- |
| Department | Salary | Status |
| Sales | 46K…50K | Senior |
| Sales | 31K…35K | Junior |
| Systems | 66K…70K | Senior |
| Marketing | 41K…45K | Junior |

age =36~40

|  |  |  |
| --- | --- | --- |
| Department | Salary | Status |
| Marketing | 46K…50K | Senior |

age =41~45

|  |  |  |
| --- | --- | --- |
| Department | Salary | Status |
| Systems | 66K…70K | Senior |

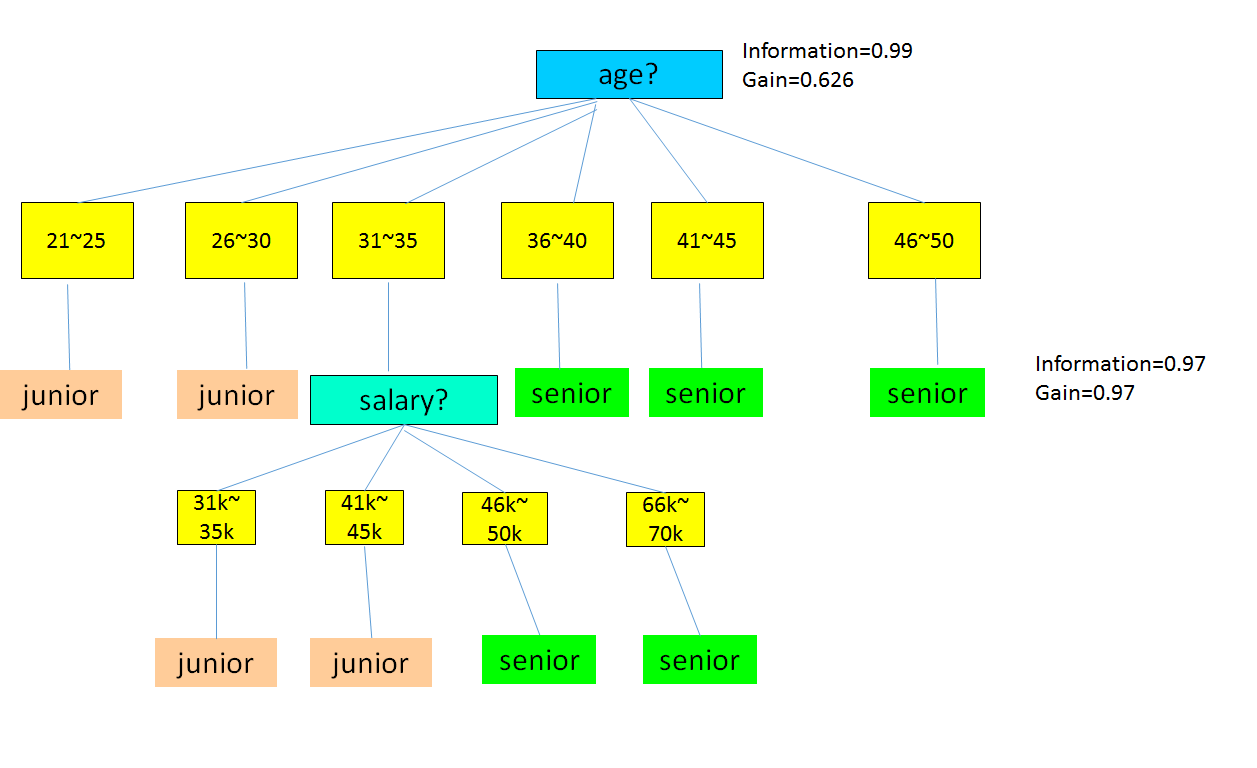
age =46~50

|  |  |  |
| --- | --- | --- |
| Department | Salary | Status |
| Secretary | 36K…40K | Senior |

|  |  |
| --- | --- |
| age= 21~25 | Pure , all status is junior |
| Age=26~30 | Pure , all status is junior |
| Age=31~35 | Info(D) = I(2,2) = 1  Info(D) of department = 0.5 Gain(department) = 0.5  Info(D) of salary =0 Gain(salary) = 1  Choose salary as second node |
| Age=36~40 | Pure , all status is senior |
| Age=41~45 | Pure , all status is senior |
| Age=46~50 | Pure , all status is senior |

1. Third level

|  |  |
| --- | --- |
| Age=31~35  Salary = 31K…35K | Pure , all status is junior |
| Age=31~35  Salary = 41K…45K | Pure , all status is junior |
| Age=31~35  Salary = 46K…50K | Pure , all status is senior |
| Age=31~35  Salary = 66K…70K | Pure , all status is senior |



1. Naive Bayes classification Using Laplacian correction

Class = junior

|  |  |
| --- | --- |
| Using Bayes therem | P(C|X) = P(X|C) \* P(C) /P(X) |
| Calculate the  P(class=junior| department=systems,  Age=26…30,  Salary=46k…50k) | 1. P(X|C) = P(X1|C) \* P(X2|C)\*…\*P(Xn|C)   = P(department=systems | class=junior)  \* P(Age=26…30| class=junior)  \* P(Salary=46k…50k | class=junior)  =2/6 \* 3/6 \* 2/6 =0.055   1. P(C) = P(class=junior) = 6/11 ~=0.55 2. P(X|C)\* P(C) =   0.055\*0.55 ~=0.03 |
| Calculate the  P(class=senior| department=systems,  Age=26…30,  Salary=46k…50k) | 1. P(X|C) = P(X1|C) \* P(X2|C)\*…\*P(Xn|C)   = P(department=systems |class = senior)  \* P(Age=26…30|class= senior)  \* P(Salary=46k…50k |class= senior)  =2/5 \* 1/10 \* 2/5~=0.016  -----   * original   P(Age=26~30|senior) =0  P(Age=31~35|senior)=2/5  P(Age=36~40|senior)=1/5  P(Age=41~45|senior)=1/5  P(Age=46~50|senior)=1/5   * **Use Laplacian correction**   P(Age=26~30|senior) =1/10  P(Age=31~35|senior)=3/10  P(Age=36~40|senior)=2/10  P(Age=41~45|senior)=2/10  P(Age=46~50|senior)=2/10   1. P(C) = 5/11 ~=0.454 2. P(X|C)\* P(C) =   0.016\*0.454 ~=0.0072 |
| Compare  P(class=junior| department=systems,  Age=26…30,  Salary=46k…50k) and  P(class=senior| department=systems,  Age=26…30,  Salary=46k…50k) | P(class=junior| department=systems,  Age=26…30,  Salary=46k…50k) is larger |